The Inadequacy of Rhythm Metrics to Quantify L2 Suprasegmental Characteristics

He, Lei

Abstract: The study investigated the L2 speech rhythm of Chinese English speakers (L1 = Mandarin) using the metrics of ΔV, ΔC, %V, VarcoV, VarcoC, rPVI-C and nPVI-V. Five native speakers of American English and Mandarin were recruited to record five sentences in English. In addition, the Chinese speakers also recorded five Mandarin sentences. One-way ANOVAs were conducted to see if significant differences exist on each of the metrics among L1 English, L2 English and L1 Mandarin. Results show that the two L1's are categorically distinct on all metrics, conforming to the perceptually distinct rhythmicities of English and Mandarin. However, no significant differences were found between L1 and L2 English (which have different intuitive rhythmicities) on almost all the metrics, suggesting that the metrics are inadequate to capture the suprasegmental details that give the final make-up of speech rhythm. Finally, new directions of speech rhythm research and new applications of the rhythm metrics are sketched.

The Inadequacy of Rhythm Metrics to Quantify L2 Suprasegmental Characteristics

Lei He

Phonetics Laboratory, University of Zurich, Switzerland

lei.he@uzh.ch

Abstract

The study investigated the L2 speech rhythm of Chinese English speakers (L1 = Mandarin) using the metrics of ΔV, ΔC, %V, VarcoV, VarcoC, rPVI-C and nPVI-V. Five native speakers of American English and Mandarin were recruited to record five sentences in English. In addition, the Chinese speakers also recorded five Mandarin sentences. One-way ANOVAs were conducted to see if significant differences exist on each of the metrics among L1 English, L2 English and L1 Mandarin. Results show that the two L1's are categorically distinct on all metrics, conforming to the perceptually distinct rhythmicities of English and Mandarin. However, no significant differences were found between L1 and L2 English (which have different intuitive rhythmicities) on almost all the metrics, suggesting that the metrics are inadequate to capture the suprasegmental details that give the final make-up of speech rhythm. Finally, new directions of speech rhythm research and new applications of the rhythm metrics are sketched.

Index Terms: rhythm metrics, inadequacy, L2 English

1. Introduction

Most human communities have certain speech styles that are constrained to fit an external or imposed periodic intervals or beats, manifesting rhythmic patterns [1, 2]. This music-like feature makes poetry and nursery rhymes possible. Apart from this artificially created artistic feature of speech rhythm, languages, in their non-artistic forms, have repetitive patterns that are at least intuitively detectable. Moreover, rhythm is among the first acquired phonological features in first language acquisition. According to [3], "during the last trimester of intrauterine development, the fetus is known to be actively processing the sound of its mother's speech." After being filtered through the amniotic fluid (analogous to a low-pass filter), the fetus can only recognize the "melody and rhythm of the language" [3: 43]. Through the non-nutritious sucking technique, researchers discovered that neonates were able to distinguish rhythmically different languages from their mother tongue [4, 5, 6]. Perceptual experiments among adults and monkeys [7, 8, 9] also yielded similar results that languages of different rhythmicities are distinguishable, whereas rhythmically similar languages are not discernable.

Early researchers [10, 11] proposed two major classes of speech rhythm, i.e. "Morse-code" and "machine-gun" rhythm, which correspond to the widely used terminologies as "stress-timed" and "syllable-timed" respectively mentioned in [12: 54] as simple rhythm units. [13] adopted this dichotomy and claimed that languages either have isochronous feet (i.e. inter-stress intervals) or isochronous syllables (i.e. inter-syllable intervals). However, later investigations on the acoustic signals, nevertheless, failed to find exact isochrony in neither inter-stress nor inter-syllable intervals, for example [14], [15]. [16] even rejected "stress-" and "syllable-timing" as metalinguistic terms, due to the failure to find true isochrony instrumentally, and claimed that the rhythmic differences between languages were the result of phonologic rule idiosyncratic to different languages.

Departing from finding absolute syllabic or foot isochrony, researchers began to delve into the structural characteristics of languages and proposed that the two types of languages have varied degrees of vowel reduction and different complexities in syllable structures: Germanic languages such as English and German tend to reduce or centralized unstressed vowels and have more complicated syllable structures, whereas Romance languages such as French and Italian normally do not have obvious vowel reductions and have less syllable weights [14, 17].

[8] quantified this idea by calculating the durational standard deviations of vocalic intervals (linear composition of adjacent vowels) and consonantal intervals (linear composition of adjacent consonants) in an utterance (ΔC and ΔV respectively), and the proportion of vocalic duration out of the whole utterance (%V). Instead of measuring the global variability of interval durations, [18] and [19] averaged the durational differences between consecutive vocalic or consonantal intervals, and called their metrics the pairwise variability indices (PVIs). Moreover, the calculation of vocalic PVI is normalized (nPVI-V) to account for tempo changes, and the raw PVI (rPVI-C) is retained to calculate consonantal PVI. [20] also normalized the speech rate by taking the ratio between ΔC (or ΔV) and the mean duration of the intervals being analyzed (VarcoC and VarcoV). These rhythm metrics have fair success in categorizing canonical "stress-timed" and "Syllable-timed" languages (Germanic languages vs. Romance languages), for example [8], [19], [20], and [21]. Moreover, the metrics have been applied in L2 prosody [22, 23], pathological speech [24], and musicology [25, 26] as well.

However, whether these metrics are robust measures of speech rhythm is strongly debated [45, 46]. Based on different elicitation methods and materials, [48] analyzed the speech rhythm of six languages and concluded that the metrics scores were easily influenced by elicitation methods and materials, and therefore, unsafe to classify languages. Also, the metrics do not have much success in distinguishing intuitively very different L2 speech from the L1 in terms of rhythm [23]. This study aims to partially replicate [23] with different speakers to further examine the robustness the metrics on L2 speech.

2. Method

2.1. Informants

Five native speakers of American English and five native speakers of Mandarin Chinese participated in the study. The native English speech data were originally part of the pathology-free data set in [24], and were made accessible to the author after passing a web-based course "Protecting Human Research Participants" with a certificate issued by The

National Institute of Health Office of Extramural Research (USA). The Mandarin speakers (all Beijing natives) were third-year English majors in a Chinese university, and therefore were deemed post-intermediate or advance English learners. The average age of the speakers were 20 at the time of their participations, and the mean onset age of English learning is 13. They received a small remuneration upon the completion of the recording. Both groups read and recorded five English sentences; besides, the Chinese group also read and recorded five Mandarin sentences. Therefore, both between-subjects and within-subjects comparisons between L1 English, L2 English and L1 Mandarin can be made.

2.2. Materials

The English sentences were the ones used in [21, 22, 24] and the average length is 16.2 syllables per sentence. Mandarin sentences were created to reflect natural syllabic distributions in daily usage, i.e. less used syllables were avoided to frequent the sentences. Similar to the English sentences, distribution of stress and unstressed syllables was uncontrolled for [21]. Glides (/w/ and /j/) and liquids (/l/) were avoided because the boundary between an approximant and a vowel is hard to discern on the spectrogram [21]. The average length is 16.6 syllables à Mandarin sentence. The annex lists all the sentences.

2.3. Apparatus and procedures

The Chinese speakers were recorded individually in a quiet room. Before the recording started, they were given adequate time to familiarize the reading materials. They were required to read sentence by sentence at normal speed. In case of stuttering, they were asked to read the problematic sentence again until totally at ease with that particular sentence. In addition, they were encouraged to reduce the number of unnecessary pauses; however, they could pause at the end of a prosodic phrase, which is normal in daily speech. They were required to read Mandarin sentences first and English sentences next. All recordings were made by the Microtrack 24/96 solid state recorder with the Audio Technica 8531 headset microphone (Sampling rate = 48 kHz, bit-depth = 16). The sound files were later transferred to the computer hard disk for further analysis.

2.4. Segmentation and measurements

The author identified and labeled the vocalic and consonantal intervals by visual inspection of waveforms and wideband spectrograms displayed in Praat [27] with the assistance of audio signals. All speech data were segmented according to the segmentation protocol set forth in [21]. The durations of vocalic and consonantal intervals were measured using a Praat script. The metrics scores were calculated on the Excel spreadsheet, and statistical testing was done using R [28].

3. Data analysis and results

3.1. Descriptive statistics and data normality

Means and standard errors of the metrics scores across L1 English, L2 English and L1 Mandarin are presented in Table 1. Both Kolmogorov-Smirnov test and Shapiro-Wilk test were employed to assess data normality. Results of both tests all indicated that the data are normally distributed (all p's > 0.05, two-tailed, see Table 2 for test statistics), meeting the normality assumption of parametric statistics.

3.2. Inferential statistics

One-way ANOVAs were conducted on all the metrics scores, and the main effect of language was found on ΔC ($F(2, 12) = 46.03$, $p < 0.0001$, adjusted $R^2 = 0.8655$), ΔV ($F(2, 12) = 11.61$, $p < 0.005$, adjusted $R^2 = 0.6026$), %V ($F(2, 12) = 15.54$, $p < 0.0005$, adjusted $R^2 = 0.6751$), VarcoC ($F(2, 12) = 11.65$, $p < 0.005$, adjusted $R^2 = 0.6033$), VarcoV ($F(2, 12) = 15.32$, $p < 0.0005$, adjusted $R^2 = 0.6716$), rPVI-C ($F(2, 12) = 41.56$, $p < 0.0001$, adjusted $R^2 = 0.8528$), and nPVI-V ($F(2, 12) = 36.05$, $p < 0.0001$, adjusted $R^2 = 0.8335$). The effect sizes (R^2) are large according to [29]'s criterion.

Tukey HSD post hoc multiple comparisons (please also see Table 1 for reference) indicated that L1 Mandarin is significantly greater than L1 English on all the metrics except %V (significance levels range from $p < 0.05$ to $p < 0.0001$), and is significantly lower than L1 English on %V ($p < 0.0005$).

Similarly, L1 Mandarin is significantly greater than L2 English on all the metrics except %V (significance levels range from $p < 0.01$ to $p < 0.0001$). On %V, L1 Mandarin is

Table 1. *Means (std. errors) of the metrics. The wave line enclosed part are not statistically different (α = 0.05)*

	ΔV	ΔC	%V	VarcoV	VarcoC	rPVI-C	nPVI-V
L1 English	46.1004 (1.40705)	56.9879 (1.87970)	41.5554 (.83937)	59.3842 (4.17688)	48.9780 (2.16890)	67.5853 (.56329)	66.7095 (2.41580)
L2 English	52.6955 (3.12098)	59.0983 (1.91801)	44.6327 (.60266)	51.7630 (1.89642)	47.3238 (2.07840)	69.1565 (3.33706)	57.4366 (1.00518)
L1 Mandarin	36.8405 (2.16055)	34.5581 (2.20080)	51.5982 (2.01049)	35.6646 (2.77137)	35.2209 (2.34678)	41.7336 (2.37952)	39.5189 (3.00890)

Table 2. *Results of the Kolmogorov-Smirnov and Shapiro-Wilk tests (two-tailed).*

		ΔC	ΔV	%V	VarcoC	VarcoV	rPVI	nPVI
Kolmogorov-Smirnov Test	D	0.830	0.379	0.792	0.538	0.558	1.022	0.707
	p	0.497	0.999	0.557	0.935	0.914	0.247	0.700
Shapiro-Wilk Test	W	0.9702	0.9660	0.9433	0.9200	0.9469	0.9324	0.9461
	p	0.8616	0.7946	0.4254	0.1926	0.4765	0.2965	0.4657

significantly lower than L2 English ($p < 0.01$).

No significant differences were found on almost all the metrics except nPVI-V (all p's > 0.1) between L1 and L2 English. Nevertheless, L1 English is significantly greater than L2 English ($p < 0.05$) on nPVI-V.

To sum up, the rhythm metrics have fair success in distinguishing canonically "stress-timed" L1 English from "syllable-timed" L1 Mandarin. However, they were insensitive to the differences between L1 and L2 English, a result quite similar to [23].

4. Discussion

4.1. Metrics scores of L1 English and L1 Mandarin

As the results suggested, L1 English is significantly higher than L1 Mandarin on ΔV, VarcoV and nPVI-V. This conforms to the fact that English have higher degrees of vowel reductions in unstressed syllables. Besides, English has phonemic distinctions between tense and lax vowels. The concomitant length differences also contribute to the higher variability in vocalic interval durations. Likewise, the proportion of vocalic duration out of the whole utterance duration is significantly lower in English than in Mandarin, also because of the occurrence of reduced vowels.

Moreover, L1 English is significantly higher than L1 Mandarin on all the consonantal metrics (ΔC, VarcoC and rPVI-C), showing a greater durational variability in consonantal intervals. Such higher variability reflects the more complicated syllable structure of English. An English syllable can be as light as V, or as heavy as CCCVCCCC; whereas even the most complicated Mandarin syllable has a simpler structure of CGVN or CGVG (N refers to the nasal; G refers to the glide, which is often acoustically realized as part of a diphthong) [31]. Such results as shown by the vocalic and consonantal metrics scores have successfully distinguished between English and Mandarin, two typical languages showing "stress-" and "syllable-timing" rhythm, agreeing with previous studies, such as [19] and [23].

4.2. Insensitivity of rhythm metrics on L2 English and critiques of the rhythm metrics

Although rhythm metrics have fair success categorizing typical languages, it fails to measure the difference between L1 and L2 English as the results of this study indicate. Intuitively, L1 English and L2 English by Mandarin speakers are rhythmically different, and [23] even claimed that Chinese L2 English was impressionistically "syllable-timed". However, L1 and L2 English are not significantly different on all the metrics except nPVI-V.

Such results suggest that the participants in the study have achieved a high level of English learning, and have already acquired such phonological aspects as vowel reductions, weak forms, and syllable structures. It would not be difficult to imagine that if our Mandarin-speaking informants were beginners of English learning, the metrics scores would have been closer to those of Mandarin, because the L1 would have still taken a substantial proportion in the interlanguage system (see [32]'s Ontogeny and Phylogeny Model of L2 phonological development that sketches the chronological trajectories of L1, L2 and language universals in the interlanguage).

Insofar as syllable structure is concerned, inexperienced learners whose L1 has simpler syllable structure always epenthesize a vowel to break down a consonant cluster or delete one or more consonants to slim down a syllable onset or coda to fit the complex L2 syllable into a legitimate one of the L1 [32]. For example, [33] discovered that the epenthesis of the schwa was common among L1 Mandarin speakers' English production (e.g., /vɪg/ → [vɪ.gə]) to conform to the syllable structure of Mandarin. This way, longer consonantal intervals are truncated, resulting in lower durational variability of the consonantal intervals. Furthermore, [34] found that experienced and inexperienced L1 speakers of Mandarin and other languages differed in their production of lax/tense vowels in that experienced learners produced more accurate distinctions between pairs of vowels like /i:/ and /ɪ/. Since the segmental length difference is often a concomitant of lax/tense distinction, inexperienced learner's speech would manifest less variability in vocalic intervals, resulting in vocalic metrics scores more similar to those of Mandarin.

That experienced L2 English learners have acquired the syllable structure and segmental length difference can easily hoax the metrics that rely solely on interval duration variability. Therefore, L2 English is classified as similar to the L1 variety, although it sounds rhythmically different from L1 English. Hence, the metrics are not sensitive enough to capture such suprasegmental characteristics of L2 English at all stages of interlanguage development, at least for Mandarin speakers as shown in this and [23]'s studies.

At the methodological level, differences in interval duration variability are not the whole story of speech rhythm, thus using the metrics as the litmus test of speech rhythm overlooks many aspects in the speech signal. Rhythm (and not just speech rhythm) is characterized by the occurrence of prominent elements at regular or semi-regular intervals. In human speech, potential cues to prominence include f0, intensity, spectral quality and duration, and languages may be different in the selection of the cues. For instance, [35] discovered that f0, intensity and duration all play a role in cueing prominence as perceived by native English speakers; however, only f0 is functional in cueing prominence in Mandarin. In a preliminary attempt to examine the difference of other cues between L1 English, L1 Mandarin and L2 English, the author analyzed the same speech data in the present study in terms of syllabic intensity (measured as dB_{SPL}) variability, and found that the difference between L1 Mandarin and L2 English was due to chance alone; however, both L2 English, L1 Mandarin were significantly different from L1 English [36], suggesting that other prosodic aspects should be included in speech rhythm research.

4.3. New directions of speech rhythm research

Apart from interval durations and intensity variability, f0 is proved effective to signal prominence or has the effect of changing the perceived duration [37, 38, 39, 40, 41]. [42] discovered that native speakers of Swiss German and Swiss French/Metropolitan French differed in the weighting of pitch cues and durational cues in perceived rhythm. [43] incorporated the language-specific weighting values of pitch and duration into combined pitch-duration PVI, and found more similar scores than otherwise would be if calculated by traditional PVIs. This suggests that perceived rhythm may not be that divergent across-linguistically if the calculation is

scaled by language-specific weightings of different acoustic cues.

Moving away from the duration-based approach to speech rhythm, [44] adopted an amplitude-based approach that examines the amplitude modulation in the speech envelope, which is modeled as a nested hierarchy with tiers representing different prosodic units, such as feet and syllables, and the hierarchy captures different metrical patterns in nursery rhymes as different phase-locked patterns between foot amplitude modulations and syllable amplitude modulations, suggesting a methodological innovation in speech rhythm research.

4.4. New applications of rhythm metrics

Although rhythm metrics are strongly debated in speech rhythm research [45, 46], they are potentially useful in the forensic milieu, because the metrics scores manifest high individual idiosyncrasies [46, 47, 48]. With explicit emphasis on forensic applications, a series of research done at the Phonetics Laboratory of Zurich University proved that rhythm metrics are useful in speaker identification [49, 50, 51, 52].

5. Conclusion

The study investigated the robustness of rhythm metrics among L1 English, L2 English and L1 Mandarin. The results indicated that L1 English and L2 English were not significantly different on almost all the metrics, although they are impressionistically dissimilar. Such results conform to [23]'s findings. The results indicate that rhythm metrics are not adequate to quantify L2 suprasegmental characteristics and speech rhythm in general. For further research, a larger sample size including L2 learners of Mandarin who speak English as L1 is also desirable. Finally, new directions of speech rhythm research and new applications of rhythm metrics were briefly introduced.

6. Annex

6.1. English sentences

1) The supermarket chain shut down because of poor management.
2) Much more money must be donated to make this department succeed.
3) In this famous coffee shop they serve the best doughnuts in town.
4) The chairman decided to pave over the shopping center garden.
5) The standards committee met this afternoon in an open meeting.

6.2. Mandarin sentences

Standard Romanization [30], phonetic transcriptions, and English translations are shown:

1) *Dàjiě jīntiān zǎochén gēn māma qù zhèjiā chāoshì mǎi jiǎozi.* /ta tɕiə tɕin tʰian tsau ʈʂʰən kən mamə tɕʰy ʈʂɤ tsia ʈʂʰau ʂɻ mai tɕiao tsɻ/
   'My sister went to the supermarket with my mom this morning to buy some dumplings.'
2) *Tā hǎoxiǎng tīng dàjiā chàng nàbù diànshìjù de zhǔtí qǔ.* /tʰa xau ɕiaŋ tʰiŋ ta tɕia ʈʂʰaŋ na pu tian ʂɻ tsy tə ʈʂu tʰi tɕy/
   'He wants to listen to the theme song of that TV show sung by everybody.'
3) *Fùjìn zhèjiā kāfēitīng mài quánshì zuìhǎo de zhīshì dàngāo.* /fu tɕin ʈʂɤ tɕia kʰa fei tʰiŋ mai tɕʰyɛn ʂɻ tsui xau tə ʈʂɻ ʂɻ tan kau/
   'The coffee shop nearby serves the best cheesecakes in town.'
4) *Xiàozhǎng juédìng jiāng xuéxiào de zúqíuchǎng chóngxīn fānxīu.* /ɕiao ʈʂaŋ tsyɛ tiŋ tɕiaŋ ɕyɛ ɕiau tsu tɕʰiəu ʈʂʰaŋ ʈʂʰuŋ ɕin fan ɕiu/
   'The schoolmaster decided to refurbish the school pitch.'
5) *Tā gēn tóngxué shuōhǎo jīntiān zǎochén zài Kěndéjī ménkǒu jiànmiàn.* /tʰa kən tʰuŋ ɕyɛ ʂuo xau tɕin tʰian tsau ʈʂən tsai kʰən tɤ tɕi mən kʰou tɕiɛn mian/
   'She and her classmates decided to meet at the KFC franchise this morning.'

Please note that the non-IPA symbols [ɻ̩] and [ɻ] represent the rhotacized and non-rhotacized non-open central unrounded apical vowels in Mandarin [53].

7. Acknowledgements

I thank Dr. Satsuki Nakai in the School of Philosophy, Psychology and Language Sciences at The University of Edinburgh for her suggestions on the study and for having helped me with the acquisition of the native American English speech data. Thanks also go to Prof. Wu Hongyun in the School of Foreign Languages at Renmin University of China for providing me with working space during data collection, and for helping me find Chinese participants.